63-3-2

ONR Contract Nonr 1834 (33)

Final Report, April 1, 1963

Lee J. Cronbach, University of Illinois, Principal Investigator

Studies of Concept Invention

In dealing with a complex environment, a person organizes his experience in some manner that is little understood by psychologists. Investigation of this "encoding" or "schematizing" process is of considerable importance.

Until 1955, concept-formation experiments had generally been restricted to the analysis of a number-right score, so that the emergence of a concept is described quantitatively and not qualitatively. Brunswik (1947) had suggested that perceptual learning be studied by examining the "criterialities" of the various cues in the stimulus-field, i.e., the dependencies linking the subject's response to the cues. This proposal enables us to elicit more information from an experiment than is provided by the conventional learning score, which counts coincidences between the subject's responses and a key. In 1955, Smedslund applied Brunswik's concepts to a task in which the subject was required to make judgments about a series of complexly varying stimulus figures, indicating what position on a numerical scale corresponded to the figure. The correct answer was calculated by the experimenter as a linear function of certain measures of the figure. Smedslund's task proved exceedingly difficult for his subjects, both because his figures were unstructured and because the feedback following each response consisted not of the correct answer but of that answer plus a random "error." As a result, cues developed criteriality for the subjects very slowly. No substantial generalizations emerged from the study, though it was an important pioneering step.

Whereas Smedslund's study was closely modeled on Brunswik's theory of probabilistic perceptual learning, Bruner and his students extended criteriality analysis to concept-formation tasks in which responses are presumed to be intellectually mediated. The study of Robert Goodnow (see Bruner, Goodnow, and Austin, 1956) presented airplane silhouettes varying in three respects. From one to three of these two-valued cues were present on any one trial, from which S was to judge whether the plane shown was friend or enemy; again, feedback was provided to allow S to learn to classify. In this study also, probabilistic considerations loomed very large, since the investigators were concerned with S's response on trials where information was incomplete.

In initiating the present study, we entertained a hypothesis roughly stated as follows: When a person encounters a series of objects (events, figures) of a certain type and observes certain associated properties or consequences, he will come to rely on certain aspects of the object as predictors. Ultimately, he will develop a "theory" about the class of objects, so that from its characteristics he can predict which of the

possible consequences will occur, i.e., he will infer the properties of the object from selected data. While this process is frequently perceptual and unverbalized, with continued experience a person should be able to develop explicitly formulated rules for the predictions. (Much of higher education is intended to facilitate just such theorizing.) A key step in this hypothesized process is identification of dimensions or constructs, i.e., the structuring of the stimulus set. It was our hypothesis that the person invents dimensions for thinking about unfamiliar stimuli objects, or, more specifically, that from the indefinitely large number of partly redundant attributes that could be used to describe differences among the stimuli, he selects certain ones that he often finds relevant to criteria, and thereafter uses these attributes as a framework in establishing predictive laws even for new criteria. It was our long-range program, therefore, to study the process by which, through a series of concept-formation tasks involving the same stimulus-family, the person develops a conceptual structure regarding that family.

We proposed to use the correlational measure of criteriality suggested by Brunswik to identify each subject's reliance on each stimulus aspect at various stages in training. In particular, we hoped that the criterialities would provide an objective indication of the "hypotheses" dominating S's response at any point, so that we could experimentally examine the extent to which verbalization helps or hinders learning. Whereas Bruner had dealt with stimulus sets described by a small number of intersecting attributes, each two- or three-valued, we proposed to use stimuli varying continuously in several dimensions. In Bruner's case, the attributes to be considered are explicitly identified and S has only to learn their relevance; we wished to study S's elicitation of attributes from a rather unstructured situation. The first step in our program was to establish that criterialities could be determined and interpreted. Our two years of experimentation led to the conclusion that the criteriality model is seriously inadequate, and that designs radically different from those employed in our studies and those of our predecessors are required to obtain information about the concept-invention process.

The chronology of our experimentation was as follows: Prior to initiation of the project (July 1960), Hiroshi Azuma gathered data (under Cronbach's direction) for a dissertation which he completed under the direction of L. M. Stolurow in summer 1960. There were four subgroups in this study. Group I was confronted with a "deterministic" situation in which the correct answer k was perfectly correlated with certain stimulus information. Group II was confronted with a probabilistic situation in which the correct answer was only probabilistically related to stimulus information. (Groups III and IV were not studied in this project.) The dissertation employed criteriality analysis, but drew conclusions only from criteriality curves representing group averages. This could not give insight into the process of hypothesis formation by individuals, and much of 1960-61 was devoted to re-analysis of these data to study individual criterialities. In the summer and fall of 1961, another experiment was run with a deterministic problem. In describing the work, we will consider first the studies under deterministic conditions, discussing experience with criteriality analysis before going on to results.

Deterministic concept formation

The criteriality model. The principal barrier to the study of concept invention proves to be the extreme difficulty of the task, even for mature subjects. Inherent in our conception (following Brunswik) is the idea that stimuli are to be encountered in random order. We found that subjects had great difficulty in holding in mind the information from successive trials; there is evidently enormous retroactive inhibition when the subject has not succeeded in organizing the stimulus set conceptually. (In one preliminary study a bright subject solved our basic circle-cross problem in five trials and gave correct responses consistently thereafter. Turning to another problem of the same sort, supposedly of the same difficulty, he was still performing at a random level after 250 trials. The process of happening on a fruitful hypothesis is manifestly unreliable in a strange situation.) The failure of subjects to learn forced us to reduce the number of degrees of freedom in the stimulus situation to four. The stimulus used throughout our studies consisted essentially of a square containing two points ... located on a grid. In the 1960 study the points were indicated as a circle and cross. In the 1961 experiment we used an arrow connecting the two points. The x and y coordinates of the points could take on values 1, 2, 3, or 4; the correct answer was a weighted average of the two x coordinates (circle or head of arrow weighted more). While it was possible for S to use many diverse cues (e.g., distance of circle from corner), we confined analysis to the criterialities of the x and y coordinates.

Brunswik's concept of a criteriality calls for computing correlations between responses and cues. It was his view that in perceptual learning one comes to recognize the validity of various cues (e.g., of surface differentiation as a cue to distance), and that the cue-response correlation (criteriality) could be compared with the correlation between cue and correct response (validity). Such a correlation would be determined over all objects in the ecological environment. While such correlations can certainly be calculated, we found that this model does not correspond well to response processes in a concept-formation task. Both the subjects' response protocols and their verbalizations indicate that hypothesis formation does not proceed through a gradual modification of cue-response contingencies. Rather, there is a discontinuous process. S forms a hypothesis about a subset of the stimuli and modifies it if it is disconfirmed. At the end of the training, he may well have a set of such hypotheses, each applying to a different subset of the stimuli; there may be additional stimuli not yet falling into any subset. One such "Type hypothesis" is the rule that "When the circle and cross fall in the same column, the number of the column gives the value of k." (This hypothesis is correct in our problem.)

The over-all criteriality gives a gross measure of cue-response correspondence, but to reproduce S's processes a two-step mathematical model with separate formulas for determining "Type" and for determining response given Type, or a complex nonlinear formula, would be required. There appears to be little possibility of actually fitting such models, because large numbers of data-points are needed to determine complex surfaces. Since, on any trial where feedback is given, S may modify his hypothesis-system, the function

to be fitted changes frequently. Instead of fitting an over-all function, one can try to account for S's responses by categorizing the stimuli a priori and examining criterialities within each subset. In our study, it made sense to subdivide stimuli according to the absolute difference between x coordinates of circle and cross, since many subjects reported rules involving these four categories as types. In the initial experiment, however, the distribution of stimuli provided relatively few examples of certain types, and did not space the types systematically over the training series. Within such types the distributions of x and x' were no longer rectangular over the range 1-4. (Particularly, for $x - x' = 3$, x and x' were never equal to 2 or 3.) The correlational index of criteriality is, like all correlations, affected by changes in the range of variables; it therefore does not serve to describe cue-response relations within types.

To cope with these difficulties, several modifications were introduced into the 1961 experiment: (1) The types defined a priori were presented to the subject with equal frequencies. (2) Series of "test" trials without feedback were introduced at several points in the training series to permit collection of numerous data-points while S was presumably in a steady state. (3) Each stimulus was repeated, the two presentations being about 16 trials apart, in order to provide evidence of "reliability." Lack of consistency implied that S had not settled upon an hypothesis. (4) The absolute distance between response and cue was employed instead of the correlational criteriality; this measures the extent to which the responses "track" a certain cue but, as we wished, is insensitive to changes in the range of the cue.

Although all these modifications appeared to be desirable, they did not eliminate our difficulties. For one thing, Ss were exposed to relatively few examples of any one type of stimulus, and therefore had difficulty settling upon good hypotheses. Reliability was frequently low. So much confusion was generated that a subject often abandoned or altered a rule for one type of problem that had been giving him success. This was presumably a consequence of his failures on the interspersed problems of other types. Even during test trials, subjects were not "in a steady state." There was some evidence that Ss changed their hypotheses even in the middle of a series with no feedback. This can be attributed either to their inability to hold hypotheses in mind or to lack of confidence.

We are forced to the conclusion that the chief aim of criteriality analysis--to trace objectively the emergence and modification of hypotheses--cannot be realized in the sort of experiment we have carried out. Various further possibilities suggest themselves. If the training series were "programed" so as to provide a regular progression of some sort, with rather frequent repetition of items, Ss would probably have greater success and their hypotheses would become more stable. In this study we held to a random sequence of stimuli in order to study a process resembling concept formation "in the real world," where members of the stimulus class are presumably encountered in random order. To introduce controlled stimulus sequences shifts us to a study of teaching as well as of learning. A second possibility is to allow S freedom in selecting stimuli, so that he can explore systematically and perhaps build up his concept by easy stages. Here, the "strategies" of the

subject instead of the decisions of a programer become an important variable. Third, one might prolong the training on the hypothesis that unreliability (instability of hypotheses) is characteristic of the first 200 trials or so, and that greater stability could be observed later.

The failure of criteriality analysis in the use to which we (like Smedslund and Bruner) put it would not have surprised Brunswik. In an unpublished 1954 paper, he discussed at length a distinction between rational and perceptual processes, hypothesizing that gradual learning of probabilities characterizes the latter while the former is marked by deliberate construction and discarding of hypotheses. Rational learning proceeds with abrupt discontinuities, rather than by imperceptibly small improvements in approximation. The distinction Brunswik makes apparently has some validity. Our findings are consistent with his expectations. Dulany finds hypothesis formation and confirmation prominent in verbal-conditioning phenomena. The "single-trial" learning reported by Estes and Bower may also conform to Brunswik's description of rational learning. To establish more clearly where performance is described by each of Brunswik's models, attention should be paid to the classification of tasks.

In passing, we may note a new possibility of "ratiomorphic" training. Dulany and O'Connell (unpublished, also Verplanck-Oskamp, unpublished) find that (i) learning with response reinforcement (such as we used) is successful for most subjects; (ii) learning where S states his rule on each trial and is reinforced if his response is right even if the rule is not is much faster; and (iii) where the reinforcement is attached to the rule learning comes much harder. But, and this is important for us, learning of type iii seems to be highly stable even under partial reinforcement (misinformation). Perhaps concept-formation studies (and inductive educational procedures) should use the type iii design.

Results. While we were not able to obtain from our design all that we had hoped for, conclusions can be drawn about group trends in this type of concept-formation task. The group averages show a steady increase in criterialities for the two relevant cues, and a decline from a low initial mean criteriality to zero for the irrelevant cues. The more heavily weighted of the two relevant cues had a higher criteriality, throughout the series of trials, than the less relevant cue. The criteriality of this less relevant cue generally developed later than that for the stronger cue.

Individual differences were remarkably consistent from the beginning to the end of the training. The correlation between accuracy score over trials 33-64 and score over trials 97-128 was .93; this is much higher than the usual interblock correlation in associative learning. Evidently, subjects who have not attained a workable (if incomplete or imprecise) hypothesis at the end of 32 trials make little progress on later trials. Verbal reports (collected in the 1961 experiment) show that poorer Ss characteristically cling to a wrong hypothesis. One S, for example, decided early to try to relate k to the length of the arrow. After 32 trials he admitted that this did not work. Nonetheless, throughout the 128 trials with feedback that followed, he clung to this basic hypothesis, elaborating it in various ways to rationalize its failures (shades of Ptolemy!). Bavelas (personal communication) finds elaboration of hypotheses rather than return to parsimonious alternatives the common response of college students to disconfirmation (which is partially reinforcing over a series of trials).

Expectancy of misinformative feedback. The 1961 experiment confirms these conclusions and deals with two new experimental variables. The first was expectancy of misinformative feedback. Noting the serious disruption of some students when their response proved wrong, we entertained the idea that providing an advance rationalization for errors might encourage them to hold to an hypothesis with which they had received some success, instead of discarding it bodily and searching wildly for a new concept. Subjects in the expectancy condition were given an explanation about our desire to simulate conditions of scientific investigation, where instrument errors and the like sometimes provide misinformation. During preliminary warm-up training, a small error was introduced into the feedback on 2 out of 10 trials; at the end of the tenth trial, this was called to Ss' attention. During the training trials themselves, feedback was invariably accurate.* The control group was given roughly similar instructions except that there was no suggestion of error in the feedback. This differs from the usual studies of misinformative feedback (and what is essentially equivalent, probability learning) in that the key variable is the subject's expectancy rather than the actual misinformation. Since, from the subject's point of view, there is no way of distinguishing misinformation from information consistent with the experimenter's predetermined concept, the variable of expectancy seems of particular psychological importance. The results showed that the group expecting erroneous feedback was significantly handicapped on the subset of problems where $x_1 = x_2$ (= k). This type of problem is much easier than the other types. On the more difficult subsets, there was no difference between groups.

The scientist, interpreting empirical data, is rendered better able to theorize by having an expectancy of misinformation (e.g., from sampling error) which allows him to tolerate irregularities in data. He raises his eyes to search for main trends, without concern for the minor anomalies in his observations. In our experiment, however, the expectation of error seems to have blinded a good many subjects to the fact that whenever circle was above or below cross in a certain column, the number of that column was the correct

---

*The device used was this. On each card was a square of masking tape. It was explained that beneath this was a number (e.g., 0.2) which represented "instrument error" such as a scientist must learn to deal with. Most of the time this number was zero, but on some trials it might depart from zero and if so, this figure was added to or subtracted from k to determine the feedback report. At the end of the pretraining trials the tape was lifted from two or three cards to show the presence or absence of such errors. There were, of course, no numbers under the tape during the main training series, though the tape was continually present as a reminder of the possibility of error.

response. In the control group 14 out of 16 learned this principle, compared with 8 out of 15 in the group expecting misinformation. In an unpublished experiment, Edwards also finds that expectancy of misinformation has a detrimental effect. Though our results suggest an interaction between this variable and concept difficulty, our design does not permit us to establish this as a finding. Over successive problems we would expect subjects to learn to cope with misinformative feedback. They must develop a data-processing strategy that emphasizes gross effects rather than trial-by-trial accuracy. The present study throws no light on whether such a strategy would emerge from continued training on many problems. One way to encourage subjects to observe trends, and not to be excessively sensitive to disconfirmations of partially correct hypotheses, may be to use block feedback rather than single-trial feedback. While "immediate" feedback usually promotes learning, it may well be that in a difficult concept-attainment task trial-by-trial feedback is more harassing than helpful. Where we wish to confirm and shape the mediating process rather than the response itself, it may be better to give block feedback, reporting a total error score for the last x trials, after every yth trial. (Perhaps, for our task, it would be suitable to let $x = 10$ and $y = 5$; this is a matter to be determined empirically.)

Stimulus diversity. The second experimental variable was the degree of continuity of the training series. In all conditions previously discussed, the coordinates x, x', y, y' took on only values 1, 2, 3, and 4, these values being marked by grid lines within the square. Stimuli of this sort are called α stimuli. The control group received only α items. The experimental group was presented with a series of mixed α and β items, β items being those where x or x' (or both) were 1.5, 2.5, or 3.5. In the present experiment, feedback was always given as a decimal fraction (e.g., if $x = 1$ and $x' = 2$, $k = 1.3$) and S was encouraged to estimate k to one decimal. It was thought that the more continuous series would produce hypotheses different in character from those formed under control conditions, in particular, that there might be a greater tendency to establish an over-all hypothesis rather than separate hypotheses for separate "types" of items.

While the training series for the experimental and control groups differed, the test series were the same. Four test series (A-D) were interspersed in the training, all of them confined to α items. The end-of-training test used a mixture of α and β items; separate scores $E(\alpha)$ and $E(\beta)$ were obtained for these two types of items. A priori, we expected the experimental group to have an advantage on $E(\beta)$ items. Neither group discussed here was led to expect misinformative feedback.

On Tests A-$E(\alpha)$ the control group trained on α items only had an advantage. Of particular interest is a comparison of $E(\alpha)$ data with a retest on the same items a week later. The experimental group was poorer on the immediate test and very slightly superior on the retest. This "sleeper effect" (of borderline significance here) is of considerable importance if confirmable. The test-retest correlation was .91 for the control group and .73 for the experimental group (diff. not significant). The gains that occurred were often quite large, and suggest that hypotheses were in some

way consolidated and simplified during the interval. On test E(β), the groups were close together--the expected advantage for the experimental group did not appear.

It is our plan to report these observations in a paper which has been partly drafted. Additional studies in the same vein are being continued in the Training Research Laboratory under the direction of Professor Stolurow. The one study completed is a replication by Thomas McHale of the 1960 study. The analysis was confined to group criteriality scores; the trends over trials confirmed the original Azuma study. Current work is dealing with similar stimuli and analytic methods, but is introducing systematic order ("programing") among stimuli in order to make the task easier.

Probabilistic feedback

In the 1960 study, a group was run under a probabilistic feedback condition. Whereas in the deterministic condition the correct answer was a linear composite of $\underline{x}$ and $\underline{x}'$ (rounded to the nearest integer), in the probabilistic condition either $\underline{x}$ or $\underline{x}'$ was equal to the correct response, $\underline{x}'$ being selected as correct on 75% of the stimuli where $\underline{x}$ and $\underline{x}'$ differed. The subject was led to believe that there was a definite rule which would give him a correct estimate of $\underline{k}$; thus the task appeared to S as a "problem-solving" rather than a "gambling" task. This study is reported in a paper in press,* and therefore, despite the significance of the results, will be reported here only in abstract.

On items where $\underline{x} = \underline{x}'$, that coordinate provides a perfectly dependable cue and most of our Ss learned to use it. There was no evidence that such highly valid cues were "undervalued" as they were in R. Goodnow's study. On the items ($\underline{x} \neq \underline{x}'$) where $\underline{k} = \underline{x}$ with $P = .75$, our more successful Ss generally made $\underline{k} = \underline{x}$ on 86-100% of the trials. This contradicts the J. Goodnow-Bruner view that "event matching" is to be expected in "problem-solving" tasks. Nor did we find the result, predicted by Bruner et. al and observed in the R. Goodnow data, of more frequent all-or-none behavior ($\underline{x}$ used as response on all trials) during the test blocks.

It appears that with probabilistic feedback and multivariate stimuli, S tends to follow the more valid cue, and that this tendency increases as training progresses. S uses any of the cues, relevant or irrelevant, to define subclasses of stimuli for which he should rely on the less valid cue. He does not seem to use hypotheses about response sequence. Criterialities are not helpful in analyzing his hypotheses because they must be calculated over a large series of trials and follow a linear model whereas the hypotheses S uses are configural.

---

*H. Azuma and L. J. Cronbach, "Concept attainment with probabilistic feedback," In K. E. Hammond (ed.), Probabilistic functionalism: Egon Brunwik's Psychology (Berkeley: University of California Press, in press).

Personnel. Dr. Hiroshi Azuma was employed as research associate on this project. He did much of the research planning and directed data collection from September 1960 to April 1962. He is now Assistant Professor of Educational Psychology at the University of Tokyo.

Donald Beane was research assistant from September 1961 to August 1962. He completed his Ph.D. in 1962 and is now Assistant Professor of Education at Wooster College.

Hiroshi Ikeda was research assistant, summer 1962. He is continuing work on his doctoral program at the Univer - of Illinois.

Lee J. Cronbach was principal investigator and took particular responsibility for data analysis and interpretation.

The Educational Testing Service, Princeton, New Jersey, rented research space to the project and provided an excellent setting for Dr. Azuma's work during the year 1960-61. The University of Illinois supported the project through the salary of the principal investigator, through facilities used in 1961-62, and through a grant from the University Research Board. Appreciation is expressed to these organizations and to the Office of Naval Research.

Appreciation is also expressed to officials of Princeton (N.J.) High School and Urbana (Ill.) High School who assisted us in obtaining subjects.